ANT GROUP

Towards More Realistic Chinese Spell Checking with New Benchmark and Specialized Expert Model

Yue Wang[1,2*], Zilong Zheng[1*], Juntao Li[1†], Zhihui Liu[2],
Jinxiong Chang[2], Qishen Zhang[2], Zhongyi Liu[2], Guannan Zhang[2], Min Zhang[1]
1 School of Computer Science and Technology, Soochow University
2 Ant Group
ywangnlp@stu.suda.edu.cn

Speaker: Yue Wang

LREC-COLING 2024

Introduction

- We propose the Realistic Chinese Spell Checking (RCSC) task. To the best of our knowledge, we are the first to investigate Chinese Spell Checking with both Chinese misspellings and pinyin errors.

- We introduce the Realistic Chinese Spell Checking Benchmark (RCSCB) and test the performance of various baselines. We find that none of the existing methods achieve satisfactory performance, highlighting the research value of our proposed task.

- We propose Pinyin-Enhanced Spell Checker (PESC), a specialized model designed to handle pinyin-related misspellings. PESC achieves state-of-the-art performance on the RCSCB benchmark while maintaining relatively low computational overhead.

In our Realistic Chinese Spell Checking (RCSC) task, considering a text sequence $X = \{x_1, x_2, x_3, x_4, \cdots, x_n\}$ consisting of n characters, wherein x_i denotes a character of Chinese, English, or pinyin, the goal of RCSC is to convert the input text X into its corresponding correct text sequences $Y = \{y_1, y_2, \cdots, y_m\}$ of m characters. Because one Chinese character usually corresponds to multiple pinyin characters, this one-to-many relationship causes that m is smaller than n.

Chinese	Pinyin	CM	PC	Chinese Text	Existing Works	Ours
✓				今天天气真不错 The weather is really nice today	✓	✓
✓		✓		今天天气镇(town)不错	✓	✓
	✓			jin tian tian qi zhen bu cuo 今天天气真不错	✓	✓
	✓		✓	jin tian tia qi zhen bu cuo	✓	✓
✓	✓			今天tian气zhen不错		✓
✓	✓	✓	✓	今天tia气镇不错		✓

Table 1: The misspelling types of Chinese text in realistic scenarios, where **CM** and **PC** denote Chinese misspellings and pinyin conversions, respectively. All the misspellings are marked in red. Existing works focus on addressing the first four misspelling types, while this work encompasses all types of misspellings.

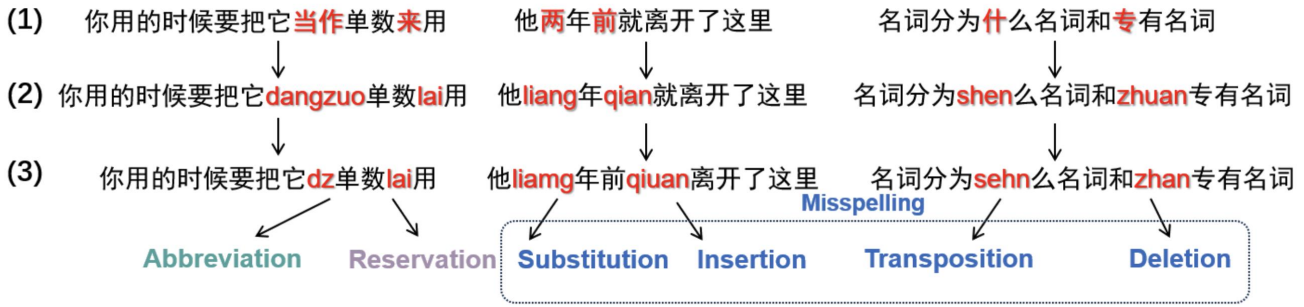

Figure 1: The illustration of pinyin conversion process. In line (1), the characters marked in red represent the Chinese characters selected to convert into pinyin; In line (2), they are converted into the corresponding perfect pinyin; In line (3), they are conducted with pinyin conversions. The pinyin conversion encompasses **Abbreviations**, **Reservations**, and **Misspellings**. The **Misspellings** include **Substitution**, **Insertion**, **Transposition**, and **Deletion**.

We collect the texts from two sources: Tomorrow Advancing Life English lecture audio (TAL) and SIGHAN benchmark. The probabilities of substitution, insertion, deletion and transposition are set to 40%, 20%, 20%, and 20%, respectively.

Dataset	Sen	Len	CM	PC
SIGHAN & Wang271K (Train)	284,201	45.7	398,749	649,427
TAL (Train)	284,136	27.3	220,823	354,125
SIGHAN15 (Test)	1100	32.7	703	1,754
TAL (Test)	12,220	27.1	9,527	15,072

Table 2: The statistics of used datasets, where **Sen** denotes the number of sentences, **Len** denotes the average length of the sentences in each dataset. **CM** and **PC** denote the number of Chinese misspellings and pinyin conversions, respectively.

Motivation:
(1) some English characters are over-corrected;
(2) some pinyin characters of a single Chinese character are predicted as several characters;
(3) some consecutive pinyin characters belonging to different Chinese characters are misinterpreted as a single character.

Input	BERT Output	Target
disturb事打扰bu要打扰伽	distur不是打扰不要打扰他	disturb是打扰不要打扰他
mind思xiayng主意	名思想主意	mind思想主意
这里要符合第一个yuanzc前面肯定后免否定人称	这里要符合第一个原则词前面肯定后面否定人称	这里要符合第一个原则前面肯定后面否定人称
翻译做第eureg	翻译做第二二个	翻译做第二个
无liq闹的无关的信息	无聊闹的无关的信息	无理取闹的无关的信息
这里应该yyi哥动词对吧	这里应该一个动词对吧	这里应该用一个动词对吧

Table 3: Examples of input, the corresponding output of BERT, and the target. Chinese misspellings or pinyin conversions, incorrect modifications, and golden modifications are marked in red, orange, and blue respectively.

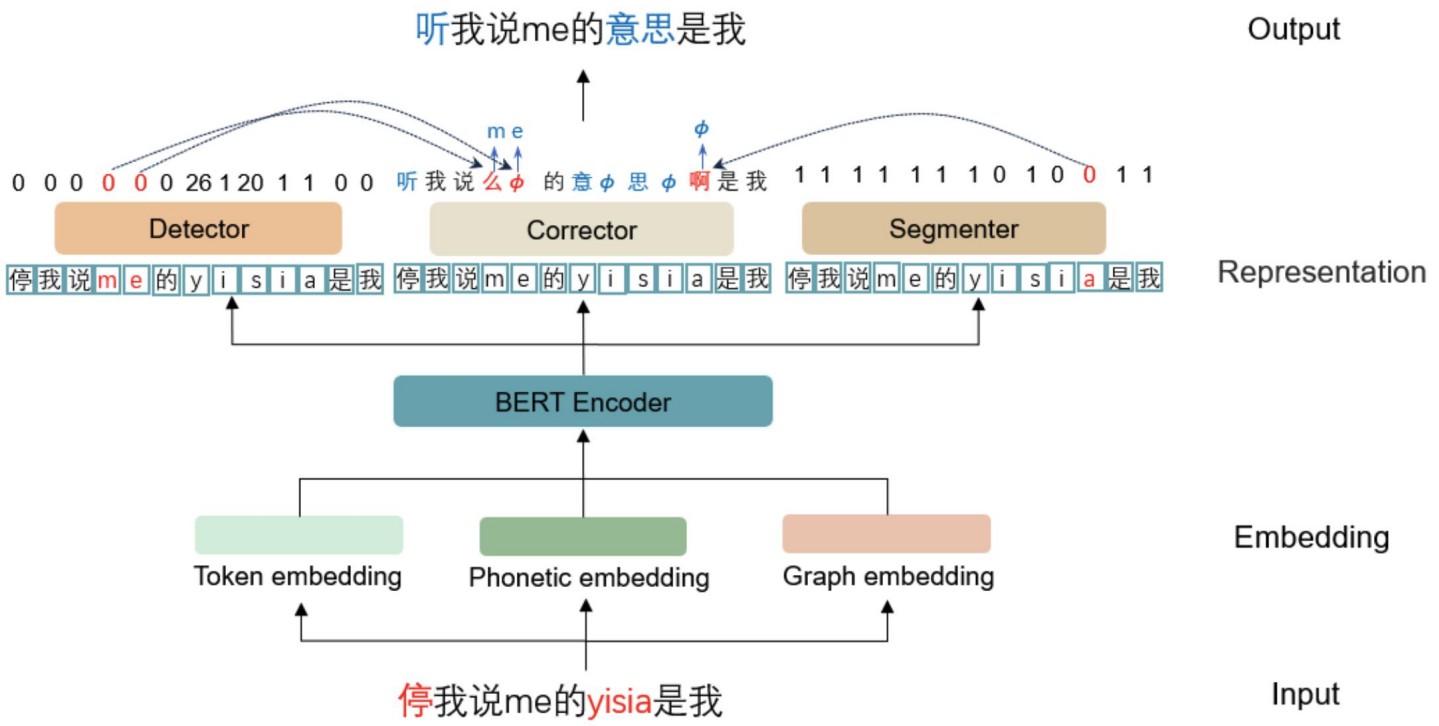

Figure 2: Overview of our model architecture. Phonetic embedding and graph embedding are used to capture phonetic and graph information. Token embedding includes segment embedding and position embedding, which is the same as BERT (Devlin et al., 2019). For example, the input sequence is “停我说me的yisia是我”. At the positions of ‘m’ and ‘e’,the corrector output are “么” and ϕ. However, the output of the detector at the same positions are all 0, which denotes that ‘m’ and ‘e’ are not pinyin characters. We abandon modifying the corrector and keep the original input "me". Meanwhile, the output of the segmenter at the position of ‘a’ is 0. We change the output of the corrector at the same place to ϕ. Finally, after deleting ϕ, the output is “听我说me的意思是我” (Listen to me, the meaning of ‘me’ is me.).

Experiments

Method	Parameters	Detection Level				Correction Level			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
T5-base (Raffel et al., 2020)	300M	-	-	-	-	77.89	71.65	71.09	71.37
Fastcorrect (Leng et al., 2021)	99M	85.82	82.15	81.44	81.79	79.25	73.28	72.65	72.96
PhVEC (Fang et al., 2022)	120M	87.86	84.74	84.35	84.55	76.29	69.17	68.85	69.01
Soft-Masked BERT (Zhang et al., 2020)	140M	79.77	74.99	73.36	74.16	69.18	60.47	59.15	59.80
BERT (Devlin et al., 2019)	120M	90.78	88.58	88.04	88.31	83.05	78.15	77.67	77.91
REALISE (Xu et al., 2021)	280M	91.31	89.43	88.67	89.05	84.57	80.31	79.63	79.96
ECOPO (Li et al., 2022)	120M	90.92	88.77	88.19	88.48	83.39	78.62	78.10	78.36
ChatGPT	175B	-	-	-	-	18.73	12.32	15.19	13.60
Yuan 1.0 (Wu et al., 2021)	12B	-	-	-	-	14.00	3.05	3.13	3.09
ChatGLM (Zeng et al., 2022)	6B	-	-	-	-	1.53	0.62	0.82	0.70
BELLE (Yunjie Ji et al., 2023)	7B	-	-	-	-	7.15	2.07	2.26	2.28
PESC(Ours)	130M	**91.85**	**89.97**	**89.43**	**89.70**	**84.92**	**80.62**	**80.13**	**80.37**

Table 5: The performance of PESC and all baseline models on the TAL test set. ***Bold*** indicates the current state-of-the-art performance. The results of LLMs are the average performance of 4 different prompts.

Experiments

Method	Parameters	Detection Level				Correction Level			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
T5-base (Raffel et al., 2020)	300M	-	-	-	-	58.67	54.46	54.86	54.66
Fastcorrect (Leng et al., 2021)	99M	74.09	71.73	72.33	72.03	67.09	63.88	64.40	64.14
PhVEC (Fang et al., 2022)	120M	69.69	67.17	69.77	68.45	57.42	53.58	55.65	54.59
Soft-Masked BERT (Zhang et al., 2020)	140M	63.82	62.83	60.40	61.59	55.09	50.43	52.87	51.62
BERT (Devlin et al., 2019)	120M	78.27	76.71	77.12	76.91	70.82	68.24	68.52	68.38
REALISE (Xu et al., 2021)	280M	79.91	78.05	79.01	78.53	72.64	69.92	70.78	70.35
ECOPO (Li et al., 2022)	120M	79.18	77.66	77.98	77.82	71.27	68.58	68.93	68.75
ChatGPT	175B	-	-	-	-	20.25	18.07	19.77	18.88
Yuan 1.0 (Wu et al., 2021)	12B	-	-	-	-	7.50	2.81	2.62	2.71
ChatGLM (Zeng et al., 2022)	6B	-	-	-	-	2.27	1.59	1.75	1.66
BELLE (Yunjie Ji et al., 2023)	7B	-	-	-	-	5.74	3.30	3.54	3.41
PESC(Ours)	130M	**81.45**	**79.90**	**80.98**	**80.44**	**73.91**	**71.34**	**72.31**	**71.82**

Table 6: The performance of PESC and all baseline models on SIGHAN15 test set. ***Bold*** indicates the current state-of-the-art performance. The results of LLMs are the average performance of 4 different prompts.

Experiments

Id	Prompt
1	一些句子会含有中文拼音，请按照示例修改给出正确的中文句子。 原句：some肯定有some对吧这个大常见了还有all就sht表示所有hamy。纠正：some肯定有some对吧这个太常见了还有all就是表示所有还有。 原句：rfan后今天下课势间也到了这ge就植能留着下次zjiang了。纠正：然后今天下课时间也到了这个就只能留着下次再讲了。 原句：**inputs**。纠正：
2	请按照示例修改掉句子中错别字和拼音，给出正确的句子。 原句：some肯定有some对吧这个大常见了还有all就sht表示所有hamy。纠正：some肯定有some对吧这个太常见了还有all就是表示所有还有。 原句：rfan后今天下课势间也到了这ge就植能留着下次zjiang了。纠正：然后今天下课时间也到了这个就只能留着下次再讲了。 原句：**inputs**。纠正：
3	请按照示例将句子中的错别字和中文拼音替换为正确的汉字。 原句：some肯定有some对吧这个大常见了还有all就sht表示所有hamy。纠正：some肯定有some对吧这个太常见了还有all就是表示所有还有。 原句：rfan后今天下课势间也到了这ge就植能留着下次zjiang了。纠正：然后今天下课时间也到了这个就只能留着下次再讲了。 原句：**inputs**。纠正：
4	请按照示例请修改错误并给出正确的中文句子。 原句：some肯定有some对吧这个大常见了还有all就sht表示所有hamy。纠正：some肯定有some对吧这个太常见了还有all就是表示所有还有。 原句：rfan后今天下课势间也到了这ge就植能留着下次zjiang了。纠正：然后今天下课时间也到了这个就只能留着下次再讲了。 原句：**inputs**。纠正：

Table 4: The different few-shot prompts used to evaluate LLMs.

Experiments

Method	Acc	Pre	Rec	F1
	Detection Level			
BERT	90.78	88.58	88.04	88.31
+Seg.	91.05	88.90	88.35	88.62
+Det.	91.12	89.02	88.43	88.72
+Det.&PED.	91.16	89.10	88.51	88.80
+ Pho.&Gra. E	91.65	89.75	89.17	89.46
PESC	**91.85**	**89.97**	**89.43**	**89.70**
	Correction Level			
BERT	83.05	78.15	77.67	77.91
+Seg.	83.16	78.25	77.76	78.00
+Det.	83.75	79.07	78.55	78.81
+Det.&PED.	83.86	79.12	78.70	78.91
+ Pho.&Gra. E	84.51	80.10	79.58	79.84
PESC	**84.92**	**80.62**	**80.13**	**80.37**

Table 7: Ablation results of the PESC model on TAL test set, where **Seg.** represents segmenter, **Dec.** represents detector, **PED.**represent pinyin-enhanced decoding, **Pho.&Gra. E** represent phonetic and graph embedding. ***Bold*** indicates the best performance.

Method	Acc	Pre	Rec	F1
	Detection Level			
Ours	**91.00**	**88.92**	**88.23**	**88.57**
All	90.57	88.34	87.75	88.04
Merge	90.96	88.82	88.23	88.52
	Correction Level			
Ours	**83.93**	**79.36**	**78.75**	**79.05**
All	82.63	77.62	77.10	77.36
Merge	83.16	78.28	77.76	78.02

Table 8: Performance of PESC (w/o phonetic and graph embedding) on the TAL test set using different alignment strategies. ***Bold*** indicates the best performance.

In the PESC model, we align the source and target sequences by associating the initial consonant with the correct Chinese character and introducing the character ϕ Additionally, we also explore other alignment strategies referred to as "All" and "Merge". "All" indicates that we associate all pinyin characters with the correct Chinese character (e.g., "jin(今)" corresponding to "今今今"). "Merge" indicates that we merge the encoder representations of all pinyin characters that correspond to a Chinese character and forward the merged representation to the corrector.

Thank You!